

Méthodes de sémantique de corpus pour la fouille de données subjectives

Mathieu Valette

travaux en collaboration avec

Egle Eensoo

Journée d'étude de l'ATALA – Fouille d'Opinions et Analyse de Sentiments – Paris (INALCO) – 21 III 2015

I
Position de la linguistique dans le TAL



TAL comme traitement automatique des langues

proposition, phrase

méthodes logico-formelles

linguistique, science du langage

TAL comme traitement automatique des textes

textes, documents, corpus

méthodes par apprentissage

linguistique de sous-traitance ?

Valette et Eensoo (2015)



linguistique pourvoyeuse de ressources

expertise linguistique ?

analyseurs morphosyntaxiques
dictionnaires, lexiques
corpus

annotations syntaxiques
annotations sémantiques

Le problème des annotations « fines »

« fines » signifie, en TAL, *lexicales.* Or les émotions ne sont pas le vocabulaire des émotions

Ex. « L'amour et la fidelité sont des espéces en voie de disparition »
= AMOUR (??)

Valette et Eensoo (2015)

L'informatique, la nouvelle science des textes ?

Maîtrise de l'appareil de production

Prolétarisation de la linguistique (et des SHS)

La linguistique doit changer d'épistémologie

Linguistique *de la langue* —> Linguistique *des textes*

Vers un paradigme rhétorique / herméneutique

Valette et Eensoo (2015)

II
La linguistique de corpus comme pré-outillage pour la fouille de textes



traitement
automatique
des langues

automatisation
des processus

visée utilitariste
performance, optimisation

reproductibilité et évaluation

linguistique
de corpus

itération
corpus < > interprète

visée épistémique
interprétation conforme

acceptabilité et consensus

Eensoo et Valette (2014ab)



Objectif

développer des modèles de fouille de textes s’adossant à la linguistique de corpus

tout en se pliant aux exigences du TAL en termes d’évaluation

Eensoo et Valette (2012, 2014ab)

Sémantique textuelle et textométrie

(i) retour au texte comme condition de l'interprétation

(ii) pas de préconception réductrice du texte

(iii) rôle du contexte global construit par le corpus de référence

(iv) fonctionnement différentiel (calculs contrastifs)

(v) développement d'une statistique contextualisante, syntagmatique ou co-occurrencielle

Mayaffre (2008), Pincemin (2010)

III
Sémantique de corpus pour la fouille de données subjectives

Etat de l'art



Approches « apprentistes »

texte comme chaîne de caractères ou de *tokens*

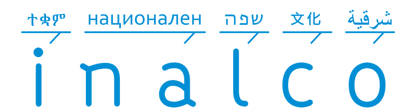
traitements massifs fonctionnant par accumulation de descripteurs

peu ou pas de sélection parmi les descripteurs

priorité donnée au choix et à l’optimisation des algorithmes

ex. Pang et al. 2002

Eensoo et Valette (2012, 2014a)



Approches « opportunistes »

enrichissement des données d'apprentissage au moyens d'éléments linguistiques rendus disponibles par des outils TAL

sélection de certains paliers analytiques
(lexiques identifiés, patrons morphosyntaxiques, parties du discours, etc.)

accent mis sur les mesures et les algorithmes

ex. Turney 2002, Yu & Hatzivassiloglou, 2003, Tanguy et al. 2011

Eensoo et Valette (2012, 2014a)



Approches cognitivistes

ressources lexicales reposant sur des modèles cognitivistes

supposition de l'existence de catégories cognitives préétablies et indépendantes des langues

unités lexicales comme des instanciations langagières d'états privés universaux

répond à l'exigence de formalisation

ex. Esuli & Sebastiani 2006, Whitelaw *et al.* 2005

Eensoo et Valette (2012, 2014a)



Approches linguistiques théoriques

revendication d'un cadre théorique
(ex : pragmatique, analyse du discours)

sélection d'éléments théoriques nécessaires à l'application

combinaison d'analyses ascendante et descendante

préférence donnée aux méthodes symboliques, mieux « contrôlables », sans exclure les algorithmes de classification

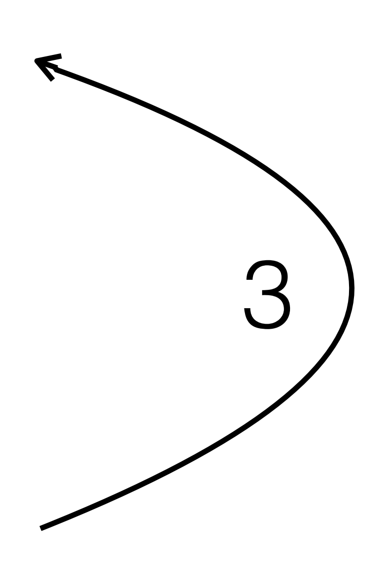
ex. Vernier *et al.* 2009

Eensoo et Valette (2012, 2014a)

Méthodologie

Méthodologie

1. identifier les descripteurs nécessaires et suffisants,
   – textométrie (Salem *et al.* 2003, Heiden *et al.* 2010),
   – calcul de spécificités (Lafon 1980)

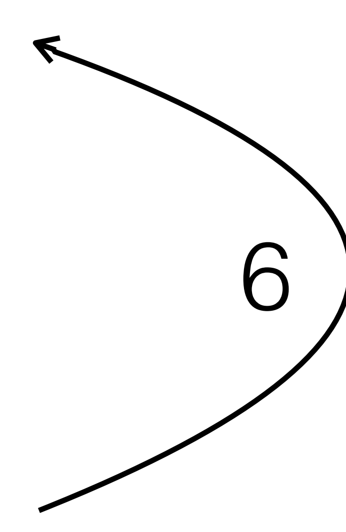




2. caractériser les descripteurs
   – sémantique textuelle (Rastier 2001)

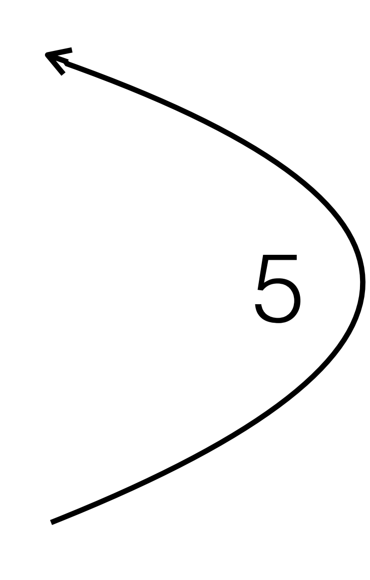
Méthodologie

1. identifier les descripteurs nécessaires et suffisants,
   – textométrie (Salem *et al.* 2003, Heiden *et al.* 2010),
   – calcul de spécificités (Lafon 1980)



2. caractériser les descripteurs
   – sémantique textuelle (Rastier 2001)

3. évaluer les descripteurs
   – Support Vector Machine (Platt, 1998)
   – Naïve Bayes Multinomial (McCallum & Nigam, 1998)



4. retour sur les qualifications sémantiques

Eensoo et Valette (2012, 2014a)

Études de cas

Étude de cas (1/2)

Sentiment Analysis

États émotionnels dans des ego-documents
du domaine sanitaire et médical

300 posts de forums de discussion (< 2012)
aufeminin.com, doctissimo.fr, etc.

Dysphorie | Euphorie

Eensoo et Valette (2012, 2014a)



Agoniste dysphorique

composante dialogique

un acteur-énonciateur égocentré (*1e pers. sing.*) et enclos sur un univers intime (*mon mari*, *ami*) :

Ce qui m'a le plus aidée ma famille, mon mari, mes enfants, mes amis

composante thématique

composante dialectique

univers impressif, non factuel, agnosique imperfectivité

Je ne sais pas comment cela va évoluer

J'ai l'impression que je vais plus mal qu'avant

Eensoo et Valette (2012, 2014a)



Agoniste dysphorique

composante dialogique

un acteur-énonciateur égocentré (*1e pers. sing.*) et enclos sur un univers intime (*mon mari*, *ami*) :

Ce qui m'a le plus aidée ma famille, mon mari, mes enfants, mes amis

composante thématique

composante dialectique

excentration de l'action, passivité

On me dit que les causes de cette maladie ne sont pas encore précises

Le médecin me dit que ça doit être le fibrome et préfère attendre l'écho

Eensoo et Valette (2012, 2014a)



Agoniste dysphorique

composante dialogique

un acteur-énonciateur égocentré (*1e pers. sing.*) et enclos sur un univers intime (*mon mari*, *ami*) :

Ce qui m'a le plus aidée ma famille, mon mari, mes enfants, mes amis

composante thématique

//diagnostic//
'syndrome', 'kg'

elle a perdu plus de 40 kg en 6 mois

//prescription//
'mg', 'chimio'

depuis février il prend 12.5 mg de cortancyl

composante dialectique

excentration de l'action, passivité

On me dit que les causes de cette maladie ne sont pas encore précises

Le médecin me dit que ça doit être le fibrome et préfère attendre l'écho

Eensoo et Valette (2012, 2014a)



Agoniste euphorique

composante dialogique

composante thématique

composante dialectique

un acteur-énonciateur altruiste (2e pers. sing.)

Alors tu vois il faut avoir espoir

Eensoo et Valette (2012, 2014a)



Agoniste euphorique

composante dialogique

qui élabore de nouveaux univers (i) en faisant part de son expérience à des fins d'édification

Pour ma part, tous c'est très bien déroulé

(ii) en intertextualisant son témoignage

Je te file une adresse : http://www.linternaute. com/ sante…

composante thématique

//médecine//

Par contre j'étais soignée à l'homéopathie, ça marchait apparemment bien

//traitement//

Elle me file un truc genre doliprane

composante dialectique

qui élabore un texte séquencé, descriptif ou argumentatif

J'ai choisi la deuxième solution, après en avoir discuté avec mon ami

Après tu t'installes puis elle va te préparer la grosse piqure mdr

Eensoo et Valette (2012, 2014a)

discrétisation

30 critères relevant de la *composante dialectique*
(représentation du temps, aspects, rôles et interactions des acteurs)

16 critères relevant de la *composante dialogique*
(positionnement énonciatif)

23 critères relevant de la *composante thématique*
17 critères domaniaux ou taxémiques (– médical)
6 critères dimensionnels (thymiques)

Eensoo et Valette (2012, 2014a)



bilan #1

TYPES	%	NB
mots simples	**68,10**	**10 700**
descripteurs dimensionnels (thymiques)	56,80	6
descripteurs domaniaux (//médical//)	61,46	17
descripteurs dialogiques	63,80	16
descripteurs dialectiques	73,09	30
dialectiques + dialogiques	**77,07**	**45**
tous les descripteurs	**84,05**	**70**

support vector machine (Platt, 1998)

Eensoo et Valette (2012, 2014a)

bilan #2

caractérisation reproductible et validée

énonciateur dysphorique
inaccompli, égocentré, clôture des univers

énonciateur euphorique
accompli, altruiste interactif, construction d’univers multiples

validation méthodologique

sémantique de corpus (analyse différentielle)
apprentissage automatique

Eensoo et Valette (2012, 2014a)



Étude de cas (2/2)

Opinion mining

Positionnement idéologique vis-à-vis de la communauté Roms
(discours médiatique)

644 commentaires d'articles de presse (2013-2014)
4 quotidiens : *Le Monde, Libération, Le Figaro, Le Parisien*

commentaires hostiles

racistes
xénophobes
défavorables distanciés

commentaires non hostiles

compassionnels
favorables distanciés

Eensoo et Valette (2014b)



Agoniste compassionel

composante dialogique	composante thématique	composante dialectique
élaboration égocentrée zone anthropique identitaire *je vs vous*	thème proximal exprimant l'empathie (*femme*, *enfants*, *misère*) thèmes distaux : opposants (*mafias, réseaux*) et adjuvants (*charité chrétienne*).	discours rapporté (guillemets, citations)

Moi, je trouve que c'est abominable d'utiliser des enfants de cette façon
(Libération, 2013-10-02)

Eensoo et Valette (2014b)



Agoniste favorable distancié

composante dialogique

anaphore (mention des *commentaires* précédents) et adresse interlocutoire (*tu*)

composante thématique

valeurs humanistes de citoyenneté (*insertion, éducation, formation*)

valeurs de respect (*racisme, haine*)

ancrage politique et sociétal (*NAM, Walls*)

composante dialectique

argumentation (*Mais, comme, comment, dont*)

Le gouvernement doit sérieusement revoir sa copie. […] Pourquoi ne pas s'attacher à une politique humaniste assise sur le programme éducation - formation - insertion ?
(Le Monde, 2013-09-26)

Eensoo et Valette (2014b)



Agoniste raciste

composante dialogique

pronoms personnels (étiquette *PRO:PER*), 1re pers. (*me*, *moi*)

composante thématique

spoliation générale : *profiter*, *argent* des *Français*, *aux frais du contribuable*, *vols*

composante dialectique

rhétorique de l'emphase (*dire que*) et de la saturation (*il y en a marre*, *encore*, *nombreux*).

Qui vole un œuf volera un bœuf c'est bien connu, si eux en sont au stade de l'apprentissage d'autres volent en toute quiétude par détournement sur les immensités du système social ! (Le Figaro, 2013-12-30)

Eensoo et Valette (2014b)



Agoniste xénophobe

composante dialogique

composante thématique

composante dialectique

retour au *pays d'origine* : *solution, renvoyer, expulser, retour, Roumanie, bulgare, dans leur pays*

politique européenne : *libre, circuler, Europe, frontière, économique*

installation en France : *s'installer, insérer, conditions, ressources*

Les européens sont libres de circuler en Europe mais ne peuvent s'installer qu'à plusieurs conditions […] ce n'est pas démanteler le camps qu'il faut mais les renvoyer dans leur pays d'origine. (Libération, 2014-01-15)

Eensoo et Valette (2014b)



Agoniste défavorable distancié

composante dialogique

posture modale de l'indigné (*je ne comprends pas*)

composante thématique

spoliation des Français par les Roms, avec la complicité de l'État (*logement*, *charge*, *payer*, *taxe*, *impôt*).

composante dialectique

narration (*depuis des années*, *puis*)

locution disjonctive (*alors que*)

ellipses (*points de suspension*),

emphase (*honteux*, *!!!!*)

Je ne comprends pas pourquoi l'ensemble des politiques laissent entendre que les roms peuvent s'installer ou ils souhaitent et bénéficier de la solidarité nationale (Libération 2013-10-02)

Eensoo et Valette (2014b)

procédure avec validation

42 critères relevant de la *composante dialectique*
(représentation du temps, aspects, rôles et interactions des acteurs)

11 critères relevant de la *composante dialogique*
(positionnement énonciatif)

90 critères relevant de la *composante thématique*

Eensoo et Valette (2014b)



validation des analyses (classification quinaire)

TYPES	Exactitude (%)	NB
mots simples (baseline)	**40**	**6 075**
lemmes	41	4 311
adjectifs et adverbes	36	878
descripteurs textométriques	**51**	**143**
descripteurs dialogiques	38	11
descripteurs dialectiques	43	42
descripteurs thématiques	47	90

Naïve Bayes Multinomial (McCallum & Nigam, 1998)

Eensoo et Valette (2014b)



validation des analyses (classification binaire)

TYPES	Exactitude (%)	NB
mots simples (baseline)	**70**	**6 075**
lemmes	72	4 311
adjectifs et adverbes	67	878
descripteurs textométriques	**77**	**143**
descripteurs dialogiques	69	11
descripteurs dialectiques	71	42
descripteurs thématiques	75	90
desc. dialogiques + dialectiques	**72**	**53**

Naïve Bayes Multinomial (McCallum & Nigam, 1998)

Eensoo et Valette (2014b)

Travaux mentionnés (1/2)

Eensoo, E. & M. Valette (2012) « Sur l'application de méthodes textométriques à la construction de critères de classification en analyse des sentiments », Georges Antoniadis, Hervé Blanchon, Gilles Sérasset, éd., *Actes de la conférence conjointe JEP-TALN-RECITAL 2012*, Volume 2 : TALN, 4-8 juin 2012, Grenoble, 367-374.

Eensoo, E. & M. Valette (2014a) « Sémantique textuelle et TAL : un exemple d'application à l'analyse des sentiments », D. Ablali, S. Badir et D. Ducard, éd., *Documents, textes, œuvres. Perspectives sémiotiques*, Rennes, Presses Universitaires de Rennes, Collection Rivages linguistiques, 75-89,

Eensoo, E. & M. Valette (2014b) « Approche textuelle pour le traitement automatique du discours évaluatif », *Études sur l'évaluation axiologique*, A. Jackiewicz, éd., *Langue française*, n° 184 (4/2014), 107-122.

Esuli, A. & Sebastiani, F. (2006). SentiWordNet: A Publicly Available Lexical Resource for Opinion Mining , *Proceedings of the 5th conference on International Language Resources and Evaluation (LREC'06).*

Hatzivassiloglou V. & Wiebe J. (2000). "Effects of adjective orientation and gradability on sentence subjectivity", Proceedings of the International Conference on Computational Linguistics, 1, 299-305

Heiden S., Mague J.-P. & Pincement B. (2010). « TXM : Une plateforme logicielle open-source pour la textométrie – conception et développement », I. C. Sergio Bolasco (éd.), *JADT 2010,* vol. 2, 1021-1032.

Lafon, P. (1980). « Sur la variabilité de la fréquence des formes dans un corpus ». Mots, 1, 127-165.

McCallum A. & Nigam K. (1998). "A Comparison of Event Models for Naive Bayes, Text Classification", *AAAI-98 Workshop on 'Learning for Text Categorization'*,41-48

Mayaffre, D. (2008), « De l'occurrence à l'isotopie. Les co-occurrences en lexicométrie », *Pour une science des textes instrumentée*, M. Valette, éd., *Syntaxe & Sémantique*, n°9.

Travaux mentionnés (2/2)

Pang, B., Lee, L. et Vaithyanathan, S. (2002). Thumbs up? Sentiment classification using machine learning techniques. Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP), pages 79-86.

Pincemin, B. (2010) - « Semántica interpretativa y textometría », in C. Duteil-Mougel & V. Cárdenas (éds), *Semántica e interpretación, Tópicos del Seminario*, 23, Enero-junio 2010, 15-55.

Platt J. (1998). "Machines using Sequential Minimal Optimization", B. Schoelkopf, C. Burges et A. Smola (eds), *Advances in Kernel Methods - Support Vector Learning*, Cambridge, MIT Press.

Rastier, F. (2001) *Arts et sciences du texte*, Paris: PUF.
Salem A., Lamalle C., Martinez W., Fleury S., Fracchiolla B., Kuncova A. & Maisondieu A. (2003). *Lexico3 – Outils de statistique textuelle, Manuel d'utilisation, Université de la Sorbonne nouvelle.*

Tanguy, L., A. Urieli, B. Calderone, N. Hathout, et F. Sajous (2011). A multitude of linguistically-rich features for authorship attribution, *Notebook for PAN at CLEF* 2011.

Turney, P. (2002) « Thumbs up or thumbs down? Semantic orientation applied to unsupervised classification of reviews », *Proceedings of the Association for Computational Linguistics (ACL)*, 417-424.

Valette, M. & E. Eensoo (2015) « Une sémantique de corpus pour la fouille de textes », *La sémantique et ses interfaces. Actes du colloque 2013 de l'Association des Sciences du Langage*, textes réunis et présentés par A. Rabatel, A. Ferrara-Léturgie et A. Léturgie, éd., Lambert-Lucas, Limoges, 205-224.

Vernier, M., Monceaux, L. et Daille, B. (2009). DEFT'09 : détection de la subjectivité et catégorisation de textes subjectifs par une approche mixte symbolique et statistique *Actes de l'atelier de clôture de la 5ème édition du Dé! Fouille de Textes*.

Whitelaw, C.; Garg, N. & Argamon, S. (2005) ACM (Ed.) « Using appraisal groups for sentiment analysis », *Proceedings of the 14th ACM international conference on Information and knowledge management*, 625-631.



Méthodes de sémantique de corpus pour la fouille de données subjectives

Mathieu Valette

travaux en collaboration avec

Egle Eensoo

Journée d'étude de l'ATALA – Fouille d'Opinions et Analyse de Sentiments – Paris (INALCO) – 21 III 2015